Sustaining Digital Research Infrastructure in the Humanities

Response to the NDRIO Call for documentation reflecting future needs

Susan Brown, Canada Research Chair in Collaborative Digital Scholarship, University of Guelph

This paper outlines some of the challenges for sustaining DRI in the humanities in relation to two connected and complementary infrastructure projects. This is not a white paper because it involves confidential information, and because it pertains to a current SSHRC Partnership Grant application.

The Canadian Writing Research Collaboratory

CWRC is a science gateway that provides an accessible onramp to digital scholarship for researchers in a wide range of humanities disciplines, particularly for using text as data. It promotes best practices for metadata and data formats, collaboration, interoperability, and preservation. CWRC provides hosting for texts, bibliographic records, and multimedia objects. Scholars can develop, analyze, and publish both research outputs and sources. **The CWRC platform offers the most open and flexible scholarly infrastructure in Canada for the production, hosting, management, sharing, and dissemination of cultural research data.**

CWRC has since its 2016 launch steadily gained both users and content. It houses 264,000 digital objects (~1.5 TB). This is a substantial amount for a humanities dataset, as it is mostly text, which is efficient to store. CWRC hosts 30 research projects such as Canada's Early Women Writers, the Orlando Project, The People and the Text - Indigenous Writing in Northern North America and Records of Early English Drama (REED) London. Last year it trained 85 HQP and had 100+ active users, i.e. content creators. As far as consumption is concerned, it saw in the past year 8173 accesses from within Canada and 3435 international accesses. It is the third-most used system on the Compute Canada cloud.

CWRC provides supports diverse modes of research. It provides individual projects with home pages to support credit and reputational practices in the humanities, as well as dashboards for managing members, workflow, reports, and project home pages. CWRC functionality includes:

Create
XML texts*
XML templates*
edit XML content in browser*
Linked Data annotation*
Named Entity Recognition (NER)
digital surrogates of texts
e-books (scanned or new)*
object metadata
bibliographic records
image, audio, video objects
thematic collections
text from images (OCR)

Read/display
custom home page
browse
search
Internet Archive Book Viewer
CWRC-Reader (XML + CSS)
Dynamic Table of Contexts
PDF viewer
image viewer
audio & video players
Voyant visualizations
Timeline/mapping views
Entity records
Credit visualization

Manage
communications
membership
roles
access*
sharing*
workflow reports
download
NER vetting/linking
backup/archiving
object versioning*
project home page
peer review processes
Traditional knowledge labels

*also openly available via Git-Writer, an open and independent version of CWRC-writer that uses GitHub for storage.

The "Digital humanities" comprise many methodologies, hence CWRC's numerous functionalities. If we consider John Unsworth's influential list of "Scholarly Primitives" (2000), CWRC supports them all: Discovering; Annotating; Comparing; Referring; Sampling; Illustrating; and Representing. CWRC thus supports a wide range of research including in sub-disciplines of **Literary Studies** including **Literary**

History, Literary Criticism, Critical Editing, and **Bibliography,** and **Canadian Literature.** Projects in **History** also include **Performance History.** A number of projects such as Canada and the Spanish Civil War contribute to **interdisciplinary** and **Canadian Studies**. We have a large umbrella project in **Indigenous** literary studies. A pending project hails from **Sociology** and **Anthropology**. Pedagogical projects have trained junior HQP while creating data on oral history, women trailblazers and COVID-19.

CWRC Sustainability

CWRC has to date been successful in sustaining itself, for which I am extremely grateful, but its position is nevertheless precarious, so I want provide some detail on what this has looked like from the inside to give a sense of the challenges of sustaining a humanities gateway. **Despite having garnered the best support available to any humanities platform in Canada since its launch, CWRC has never been able to see more than 24 months ahead in terms of its sustainability horizon.**

CWRC was created with public funds, save for modest contributions from corporate partners. Use of the platform is free for users and content is required to be Open Access where possible. User fee models would be problematic to impose the humanities, for cultural and practical reasons. Projects with bespoke Islandora front-ends (Drupal multi-sites) are responsible for their own sustainability costs, and when new projects are coming on board and applying for grant funding at the same time we ask them to budget modest amounts towards the costs of onboarding, customizing, and ingesting data. However, it has been rare for projects to be able to garner dedicated funding such as this. SSHRC review panels are not used to seeing these sorts of expenses in grant applications, can have difficulty evaluating them if there is not a digital humanist on the committee, and have been known to cut technical expenses from budgets; applicants are as a result reluctant to include them and may be disadvantaged by them. I have myself had a technical line cut wholesale from a SSHRC budget for the funding of a digital project.

CWRC has been sustained since 2016 by:

- CFI IOF funds, most of which were consumed to get the platform fully operational
- CRC funds and an associated JELF CFI; I am fortunate that Guelph agreed to generous terms for my CRC, including some time for programming; not all CRCs could do this
- 3 research software grants from CANARIE, 2 of which included operations funding
- Significant funding from a partnership with Bucknell University, which has received two grants from the Mellon Foundation to its own instance of CWRC as a digital publishing platform

Retaining personnel is essential and a huge challenge within university structures when a project is soft-funded. I am hugely fortunate that CWRC's core personnel have been astonishingly loyal. Nevertheless, it is in large part luck that they have been retained. Figuring out a means of providing project managers and developers for infrastructure projects with at least longer-term contracts would make a huge difference to stability.

Retaining key personnel is essential to the operations and maintenance of research software, because it provides core stability while other HQP, i.e. students, gain expertise and through other roles. This has been essential to CWRC systems administration, and to software maintenance and sustainability. **While some view research software development as generic, that has not been our experience with CWRC**: it took an experienced Drupal programmer a year to get up to speed on a portion of its unusual data structures and software framework, despite being based on a leading open source Drupal framework. CWRC could not survive with short-term developer time or a succession of inexperienced project managers.

Key challenges and observations related to CWRC sustainability:

- Created major job-related stress for our lead programmer. When CFI-IOF funding was exhausted, I could not retain him full-time. CWRC was saved by the U of Alberta Libraries, a stalwart partner of CWRC, who agreed to hire him for half of his time, but HR regulations at U of A meant that he had to be formally "laid off" (a painful process) and rehired.
- Retained the project manager only because she is motivated by the interest of the work and dedicated to the project, and frankly because she is relatively isolated in terms of employment and selective about where she has applied for other positions, which she has. Her job ought to be re-evaluated because she is paid considerably less than she deserves based on her credentials, expertise, and responsibility, but given the precarity of the funding, I have not been in a position to address this inequity.
- Had either of these two employees left, CWRC would be in serious jeopardy. This is in part due to the nature of its technology stack, so I am trying to design CWRC 2.0 and LINCS to be less dependent upon individuals, but it will still be complex and require dedicated management.
- Lost one senior part-time programmer after many years because shifts in funding from CFI to CANARIE, which precluded hiring independent contractors, resulted in significant cuts to his income. It dropped to about 50% of market rate, in part due to an inability to get university HR departments to band software developers appropriately, a pervasive problem when it comes to research software personnel.
- The burden on my time as project director of preparing at least two major funding applications per year cuts into my ability to engage in outreach. As a humanities CRC I am teaching the same load as many regular STEM colleagues. My research productivity has suffered as a result.

CWRC had begun discussions in 2019-2020 towards a consortium for support when the pandemic hit.

Linked Infrastructure for Networked Cultural Scholarship

The LINCS CFI-funded Cyberinfrastructure project, started in January 2020, uses Semantic Web / Linked Open Data (LOD) technology to mobilize existing humanities datasets from multiple disciplines spanning **literary studies**, **history** including **book history, literary history and art history, music, Indigenous studies, communication studies,** and **women's studies** and make them interoperable. It will also advance **digital humanities, information studies,** and **computer science** research. **Canadian universities, research libraries,** and **memory institutions** are collaborating to:

- convert data from 50+ researchers from across the humanities into LOD
- Establish a platform for accessing and using the resulting data effectively
- provide tools for continuing to create and convert data, including via an updated CWRC 2.0.

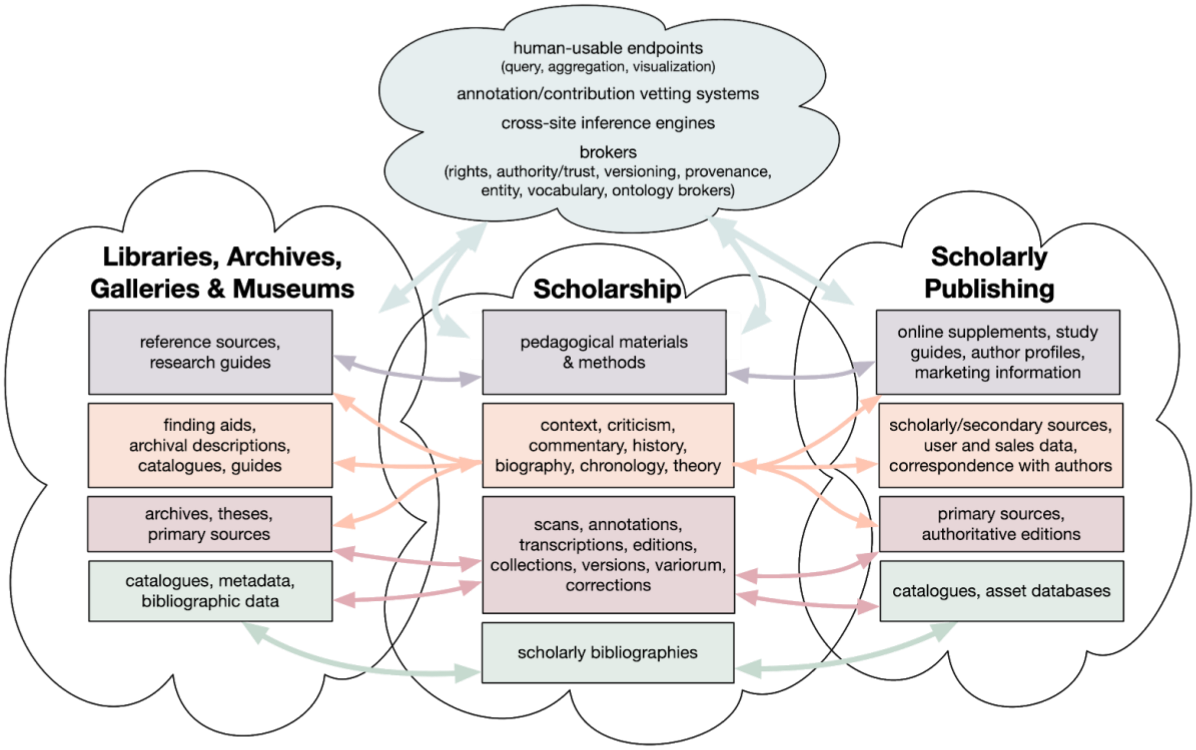
LINCS combines technical sophistication, standardization, and customizability with the promise of more legible data structures and relationships. LINCS is positioning Canada as a leader in Linked Open Data for cultural research. Despite ramping up during the pandemic it is making good progress, leveraging open-source software and platforms, creating a large linked dataset or knowledge graph tailored to the needs of cultural researchers, and developing a cohort of staff to support its use.

LINCS sustainability

LINCS builds on past crucial but tentative and disconnected efforts with LOD in Canada. With other stakeholders LINCS is collaborating on **the proposed Linked Open Data for Diversity and Difference (LOD^3) SSHRC Partnership Grant** that will, if funded, work to develop a collaborative model for a sustainable national cultural LOD infrastructure.

LOD^3 is s a multi-stakeholder, multi-sectoral partnership comprised of researchers and universities; libraries (research libraries; LAC/BAC); knowledge-sector non-profits (CARL bridging to FRDR/Portage; Scholars portal; CRKN/Canadiana/ORCID), non-profit GLAM institutions (university galleries; British Museum); government agencies (Canadian Heritage Information network); and university presses. There Is also potential for broader expansion into small/commercial presses and corporate arts sectors.

These partners belong to **an emergent LOD ecosystem within which shared infrastructure would support linkages between related content across stakeholder groups**. LOD is a "heavy" technology stack that requires dedicated expertise to establish and maintain. For that reason, none of these sectors has implemented LOD technology full. Yet its potential for research, cultural, social and economic benefits is huge.


human-usable endpoints
(query, aggregation, visualization)
annotation/contribution vetting systems
cross-site inference engines
brokers
(rights, authority/trust, versioning, provenance, entity, vocabulary, ontology brokers)
Libraries, Archives, Galleries & Museums
reference sources, research guides
finding aids, archival descriptions, catalogues, guides
archives, theses, primary sources
catalogues, metadata, bibliographic data
Scholarship
pedagogical materials & methods
context, criticism, commentary, history, biography, chronology, theory
scans, annotations, transcriptions, editions, collections, versions, variorum, corrections
scholarly bibliographies
Scholarly Publishing
online supplements, study guides, author profiles, marketing information
scholarly/secondary sources, user and sales data, correspondence with authors
primary sources, authoritative editions
catalogues, asset databases

Multi-stakeholder ecosystem for cultural Linked Open Data

The partnership aims to establish a strong network of partners that knits the scholarly community into closer collaboration with publishers, GLAM institutions, and information stakeholders. It will work on the gap analysis, policy, tools, training, and governance essential to developing a national linked data infrastructure for the social sciences and humanities. **What LOD^3 can achieve in terms of establishing a viable model for sustainable infrastructure for this ecosystem will finally depend upon the flexibility of the NDRIO model and its ability to foster multi-sector investments and partnerships.**

Key challenges and observations:

- Multi-institutional collaboration allows a combination of expertise for infrastructure building that goes beyond what any single institution can provide, and it engages stakeholders from the outset to build relationships and prepare for sustainability. The fact that the Cyberinfrastructure program didn't count against CFI envelopes hugely contributed to institutional participation.
- CWRC learned a lot from the CANARIE model of RS development enforced by their proposal and grant management process. At the same time, the available technologies for RS development advanced, so LINCS (including CWRC 2.0) will proceed on much surer footing re: sustainability.

Summary points regarding science gateways in non-traditional fields

- Uptake takes time, due to factors including lack of alignment with/scarcity of research funding in the humanities, and lack of normalization of digital research methods, which increases onboarding time. Humanities infrastructures may therefore take longer to reach levels of use expected by major operational funding streams or necessary to form a consortium.
- Sustained inter-institutional partnerships are difficult to establish and maintain. A framework for facilitating infrastructure-related consortia would be very useful.
- DRI use is in many cases quite diffuse: researchers draw from digital tools and resources across the internet, seldom acknowledge them, and will often, for instance, reference them as if they had consulted the physical resource.
- Impact is harder to assess: citation metrics are not reliable to the extent that they are in STEM, and given the diffuse usage it is hard to reach the research community that has benefitted from infrastructure .
- Mentoring infrastructure projects at early stages in both research software development best practices and management processes will be necessary to ensuring that less well-funded research communities have equitable access to NDRIO support.
- Precarity of key personnel positions is a major challenge. Short-term contracts are demoralizing. Staff often leave for less interesting but stable jobs. Their departure can jeopardize operations, but dependence upon key, underpaid, precariously employed expert personnel is typical rather than unusual for humanities research infrastructure.
- A higher proportion of staff time goes to outreach, training, and support for humanities DRI gateways, since individual research projects are less likely to be able to provide such training than e.g. a STEM lab.
- The need for GUI support for the most humanities researchers increases development, maintenance, and upgrade costs for humanities DRI platforms and gateways.
- Research software development and maintenance expertise may be less generic in the humanities than that for STEM-oriented fields.
- Greater availability of funding for new features rather than maintenance makes it harder to address problems that impede usability and uptake.
- The vicissitudes of funding make provisions for emergency or bridge funding highly desirable, since a break in operations means losing key personnel and risks the loss of the entire investment in DRI.
- The scarcity and precarity of funding for humanities infrastructure maintenance and operations in Canada has a negative impact on the quality, productivity, and uptake of the infrastructure.

I hope to have provided here some insight into the challenges for infrastructure sustainability in the humanities. I am very proud of what we have been able to do with CWRC, and of what we are doing better with LINCS and CWRC 2.0 as a consequence of what we have learned. At the same time, I am haunted by the sense of how much better CWRC would be, and how much better it could already have served the research community, had it not been such a struggle to sustain it from the very outset. I am equally haunted by the prospect of how much will be lost in public funding and human effort if CWRC and LINCS are not sustained to the point where we can truly test their potential to reshape and advance research in the humanities.